In our analysis of the National Health and Nutrition Examination Survey, we analyzed the five variables measured on each subject of the survey, age, body weight (lbs), height (in), average diastolic blood pressure, and serum cholesterol. Using blood pressure rank - categorized into three groups (one, two, and three) - as the grouping variable, we performed discriminant and classification analysis.

In section B we found that while the ages of subjects from the survey range from 20 to 90, most were around the ages of 35 and 65 seen by the two bimodal peaks. The average weight across subjects was approximately 167 pounds and the average height around 67 inches. Interestingly, the distribution of the average diastolic blood pressure was roughly bell shaped and centered around 74.91 mm Hg. The average serum cholesterol levels were around 205.63 mg/100ml

Prior to creating the correlation matrix and correlation plots, we believed that height and weight, age and weight, age and average diastolic blood, and weight and serum cholesterol, may have moderate correlations. In section C1 we were surprised to see relatively moderate to weak correlations between each of the five variables from the dataset. The correlation between height and weight was the highest of all the correlations at 0.48, thus we omitted the height variable from further analysis in sections C2 and C3. The next highest correlation was between age and serum cholesterol level at 0.28. Here, since the correlation was weak, we did not omit either variable. Correlations coefficients between all other variables were smaller than 0.2 and greater than -0.1, which we classified as weak.

In C2, the standardized coefficients from the first linear discriminant function indicated that age and average diastolic blood pressure were the most important variables in separating the blood pressure rank groups. The second linear discriminant function also indicated these two variables were most important. After conducting a significance test for discriminant functions we found that only the first function was statistically significant at the adjusted significance level of $\alpha = 0.025$. The partial F-test conducted at $\alpha = 0.0125$ reflected the same ordering of the variables from most important to least important - age, average diastolic blood pressure, body weight, and serum cholesterol - as seen from the coefficients in the first linear discriminant function. Since the p-value for serum cholesterol was greater than 0.0125, we concluded that it was not a significant contributor. Furthermore, the discriminant plot supported our analysis as it showed clearer separation between blood pressure rank groups from the first linear discriminant function than the second function.

After creating the linear classification functions for each of the three blood pressure rank groups, we were able to correctly classify the first subject into group one. While the function for group one produced the highest score, we noted that all three functions produced similar values. This was supported by our confusion matrix and accuracy measures. Our apparent correct classification rate was 63.64%, suggesting a moderate number of misclassifications.

One challenge was determining whether the correlation of 0.48 between height and weight was strong enough to omit height from the analysis to avoid multicollinearity. Ultimately, we decided

to omit the variable out of an abundance of caution. If we were to do this project again, we could have done one set of analyses with the height variable and one without to compare results. Additionally, a variable such as cortisol level could have been useful in group separation. Since stress is a big factor in blood pressure and can be measured through cortisol level, it would have been interesting to conduct the analysis with this variable and see how group separation reacted as a result.